Current challenges in clinimetrics

de Vet, H.C.W.; Terwee, C.B.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2003

DOI (link to publisher)
10.1016/j.jclinepi.2003.08.012

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 1137–1141

Journal of Clinical Epidemiology

VARIANCE AND DISSENT

Presentation

Current challenges in clinimetrics

Henrica C.W. de Vet*, Caroline B. Terwee, Lex M. Bouter

Institute for Research in Extramural Medicine, VU University Medical Center, Van der Boechorststraat 7 1081 BT, Amsterdam, The Netherlands

Accepted 15 August 2003

Abstract

Clinimetrics is a methodologic discipline that focuses on the quality of clinical measurements, for example, diagnostic characteristics and disease outcomes. Different clinimetric properties, such as reproducibility and responsiveness, are important in both the development and the evaluation of measurement instruments. This article presents a number of the current challenges in clinimetrics: there is much confusion with regard to terminology, clinimetric properties are population and situation-dependent, and the abundance of different measurement instruments in specific fields hampers the comparison of study results. Further challenges lie in the improvement of the quality of both the measurement instruments and the performance of the actual measurements, and the assessment of the suitability for use in clinical practice. From the perspective of evidence-based medicine, it is essential to have measurement instruments that make it possible to detect clinically relevant improvements that are due to diagnostic and therapeutic interventions. Close collaboration between clinicians, statisticians, epidemiologists, and psychologists is necessary to guarantee healthy future developments in clinimetrics, serving the needs of both clinical research and clinical practice.

1. Introduction

Alvan Feinstein introduced the term "clinimetrics" in the medical literature in the mid-eighties of the last century, as a methodologic discipline focussing on measurement issues in clinical medicine [1,2]. Clinical phenomena to be measured include symptoms, pathophysiologic findings, and disease status or severity. Despite the new term, the field of clinimetrics is not new. It can be considered as a branch of biometrics, a long-standing discipline of the methodology of measuring biologic phenomena. Another related metric discipline is psychometrics, which is concerned with psychologic phenomena. With the introduction of the term clinimetrics, Feinstein drew attention to the importance of specific measurement problems in medicine [2]. In this article we will describe a number of important challenges for the clinimetric research agenda in the coming years.

2. Scope

Clinimetrics focusses on the quality of clinical measurement. Quality of measurement includes both the quality of the measurement instruments and the quality of performance of the actual measurements [2]. Clinical measurement instruments encompass not only x-rays and other imaging techniques, histologic examinations, clinical chemistry measurements in serum and urine samples, questionnaires and interviews, but also patient history and physical examinations performed by care providers. The quality of performance of the measurements depends, for example, on the expertise of the persons carrying out the measurements, the quality of the sample, or the amount of attention paid by a patient to a questionnaire.

Evaluation of the quality of measurement instruments is the central issue in choosing or designing an instrument, while the quality of performance is crucial when using measurement instruments in research or clinical practice. To choose the best available measurement instrument either for research purposes or in clinical practice, one has to judge the clinimetric properties of candidate instruments from the literature or carry out empirical research to assess these properties. If the instrument we need is not yet available, a new instrument has to be developed and, of course, to be evaluated on its clinimetric properties. Investigators often prefer to develop a new instrument for their own research. However, before developing a new instrument, one should always search intensively for any available instruments that may suit the purpose at issue. It often is underestimated how many potentially suitable measurement instruments already exist, and also how long it takes to develop and evaluate a new instrument.

Both in the choice of existing measurement instruments and in the development of new instruments the aim of the

* Corresponding author. Tel. +31-20-444-8176; fax +31-20-444-8181.
E-mail address: HCW.deVet@vumc.nl (H.C.W. de Vet).

0895-4356/03/$ – see front matter
doi: 10.1016/j.jclinepi.2003.08.012

measurements plays an important role. The aim concerns the constructs or aspects one wants to measure, as well as the purpose of the measurement. One has to decide, for instance, whether the instrument is intended to discriminate between subjects at one point in time (discriminative measurement), or to measure change over time (evaluative measurement) [3,4]. This purpose determines which clinimetric properties of an instrument are most important. For example, items that are very helpful in the cross-sectional discrimination of subjects may be less useful to detect clinically relevant change [3,4]. The variable age may be included in a diagnostic index to differentiate between knee arthrosis and arthritis, but age is useless in an evaluative measurement instrument because it cannot be influenced.

For the development of questionnaires the necessary initial steps consist of item selection and item reduction. Scoring options and weighing of the items are other issues that need to be settled [2,5]. These developmental issues are not only applicable to questionnaires but also to other measurement instruments. Suppose that a new imaging technique produces an image that makes more anatomic structures or pathophysiologic processes visible than any other technique. We then have to determine which of the new elements in the images are clinically relevant for the diagnosis, prognosis, or choice of treatment, and what the most appropriate scoring categories would be.

After the development of a measurement instrument, the quality of the instrument has to be assessed [2,5]. Validity is the essential issue in the quality of a measurement instrument. A measurement is valid if it measures what it is intended to measure. Reproducibility is the extent to which repeated measurements yield the same outcome. A poorly reproducible measure decreases validity if measured only once, because it gives different outcomes at each measurement. It requires averaging of repeated measurements to diminish measurement error. For evaluative measurements, the essential clinimetric property is responsiveness [6], which means that an evaluative measurement instrument should be able to detect clinically relevant changes in health status over time. Responsiveness can be seen as longitudinal validity, meaning that an instrument should detect changes in persons who actually change and no or only small changes in persons who remain stable over time [7].

2.1. Challenges

Clinimetrics, as a methodologic discipline, offers many exciting challenges. Below we list five important challenges for the clinimetric research agenda in the coming years.

2.1.1. Confusion in terminology

There are many types of validity. Following the definitions of Streiner and Norman, criterion validity is the most powerful type. In this case a gold standard is available, and it is possible to examine the extent to which a measurement instrument provides the same results as the gold standard. In the absence of an acceptable gold standard, construct validity is the next best option. Then, for instance, the correlation of the measurement instrument under study with other instruments that claim to measure the same construct is assessed. To examine construct validity it is necessary to define hypotheses about how the scores on the instrument under study will correlate with the scores on other instruments. Confirmation of these hypotheses gives support to the validity of a measurement instrument. Content validity concerns judgment whether all important components of the construct to be measured are covered by the instrument. Face validity implies an overall judgment of adequacy on the face of it, without paying close attention to the component parts. Although, in the opinion of some, at the bottom in the ranking of powerfulness, face validity may be very important in some cases, particularly for indexes that are intended to reflect observations and intuitions of clinical experience. Unfortunately, the above definitions of criterion and construct validity are by no means universal, and many other definitions of the various types of validity also can be found in the literature [5].

Reproducibility includes two concepts: reliability and agreement [8], but the distinction between these terms is not always appreciated. Agreement represents lack of measurement error. Reliability represents the extent to which individuals can be distinguished from each other, despite measurement errors. A reliability coefficient relates the variation due to measurement error to the variation between individuals within a population. Acceptable reliability will vary depending on the circumstance. Let's take the example of body weight. It is found that repeated measurements with a weighing scale vary around the "true" weight by 0.5 kg. This is an indication of agreement. This would be acceptable reliability if the measurements are focussed on an adult population, but not when it is used to weigh babies. In a heterogeneous population, individuals are more easily distinguished from each other, and the reliability of a measurement instrument will be greater, given the same measurement error. Reliability parameters are important when the aim is to discriminate between individuals, and agreement parameters are important when one wants to detect changes in health status over time [4,7].

Responsiveness is an important clinimetric parameter for measurement instruments that aim to measure change over time, for example, outcome measures in studies on the effects of treatment [3,4]. In a short period of time, 25 definitions and 31 different formulae have been developed to calculate responsiveness [9]. These different measures can be grouped according to different conceptualizations of responsiveness [9–12]. The most important distinction is made between responsiveness measures that quantify the treatment effect (effect size) and measures that focus on the longitudinal construct validity by assessing the correlation of change scores with another measure (external standard) for change [9,12]. Norman et al. label these two categories as distribution-based and anchor-based measurements of change, respectively [12]. Suppose that in a trial on exercise therapy for

low back pain the outcome measure functional status is assessed by different measurement instruments, for example, the patients complete the Roland Disability Questionnaire and Oswestry Disability Questionnaire. Additionally the physiotherapist makes a judgement about function after physical examination. In the distribution-based approach, the instrument that shows the largest improvement in function is considered to be the most responsive one. With the anchor-based approach, the improvement on each measurement instrument is correlated with an external standard for change. This external standard can, for example, be the opinion of the patient or doctor about whether or not the health status has improved. The measurement instrument that shows the highest correlation with the external standard is considered to be the most responsive. Obviously different instruments can be identified as being most responsive by each approach. To cope with the current confusion in terminology, specification of the type of validity, including the reproducibility and responsiveness at issue, is of ultimate importance. Hopefully, over the next few years, consensus can be reached on an unequivocal taxonomy of clinimetric properties and their most appropriate operationalizations.

2.1.2. Clinimetric properties are relative concepts

Unfortunately, validity is not a "black or white" issue: being valid or not valid. First, the assessment of criterion validity is only possible if an acceptable gold standard exists. In other cases one has to rely on construct validity and to define specific hypotheses with respect to the extent to which the instrument under study correlates with other instruments (partly) measuring that same construct. These hypotheses will be confirmed or rejected, sometimes very convincingly, but in other situations the conclusions will be doubtful. If the Pearson correlation coefficient between two measurement instruments was hypothesized to be higher than 0.70, a value of 0.84 is much more convincing than a value of 0.72. We obviously have to deal with various degrees of validity, and the challenge is to decide whether the degree of validity is sufficient. Another reason why validity, reproducibility, and responsiveness are relative concepts is because these properties are situation-dependent. If a study concludes that a specific measurement instrument appears to be valid, this fact is readily adopted by researchers who need a validated instrument. However, the question of whether the instrument is also sufficiently valid for another purpose or in another situation is typically ignored. Validity is dependent on the population in which the instrument is intended to be used and on the size of the effects that the instrument is required to detect. Is the Roland Disability Questionnaire equally valid or responsive in patients with severe back pain as with mild back pain, or in patients with lumbar radicular syndrome who may also have pain in the leg? Are there ceiling or floor effects in the instrument? Furthermore, instruments that are valid for discrimination may be invalid for evaluation. Thus, clinimetric properties are very much situation-specific, and depend highly on the study population and the measurement circumstances.

Another important issue is the distinction between the validity of the measurement instrument and the actual performance of the measurement. If the measurement is performed suboptimally, the instrument itself may be sufficiently valid but the performance may not. For example, assessment in a study of the validity of biopsies in routine clinical practice should not be performed by experienced experts, but by pathologists who are a sample of the pathologists who assess the biopsies in routine clinical practice. This all illustrates that one has to look carefully at the circumstances of validity studies and be cautious in generalizing the results to other situations.

2.1.3. There are too many measurement instruments

The tendency of researchers to develop their own measurement instruments has led to an enormous redundancy in instruments. In 1987, Feinstein already found 230 instruments to measure mobility [2]. To be valuable for further use in other studies, it must be clear for which purpose the instrument has been developed and for which populations and situations it has been validated. A systematic review of the clinimetric properties of the available measurement instruments in a specific field may facilitate the choice of an instrument, that is, these reviews of measurement instruments indicate the number of existing instruments to measure the construct at issue, summarize their clinimetric properties and purposes, and give direction to further validation studies, if necessary. Good examples include those by Bialocerkowski et al. [13], who inventoried all wrist outcome instruments and described their content and methodologic quality, and Coons et al. [14], who examined the clinimetric properties of seven broadly used general Health Related Quality of Life scales (HR-QOL). We strongly feel that systematic reviews of available measurement instruments should precede the decision to develop a new measurement instrument, analogous to the need for a systematic review of available trials in a specific field before one decides to design a new trial.

Another strategy that can be applied to cope with the existence of many measurement instruments for the same construct (e.g., functional status in patients with low back pain) is to examine whether they can be compared to each other in such a way that the score on one instrument can be converted into a score on another instrument. It would, for example, be helpful to know which score on the Oswestry Disability Questionnaire corresponds with a value of 14 on the Roland Disability Questionnaire, and how many points improvement on the Oswestry Disability Questionnaire corresponds to an improvement of five points on the Roland Disability Questionnaire. Appropriate methods are still under development [15]. When many different measurement instruments to measure functional status in low back pain are used in different trials, this hampers comparison of the results. Therefore, equalization of outcome measurements would be very helpful in systematic reviews of effectiveness.

2.1.4. Improvement of measurements

The previous paragraphs focussed on the development and the evaluation of the quality of measurement instruments. The ultimate challenge for clinimetrics is to improve the validity of measurements, either by increasing the quality of the measurement instruments as such, or by improving the quality of the performance in practice. The optimal starting point is the choice of the instrument that is the most valid for the situation at issue. But further improvement of the quality of the actual measurements can often be achieved by optimizing the circumstances of performance of measurement. For example, to improve the reproducibility a number of strategies are available: training the persons who are to perform the measurements, standardization of the tests and the test circumstances, feedback to those who have shown a low interobserver reproducibility, or repeated measurements and averaging the results if the intraobserver or interobserver reproducibility remains low. This underlines the importance of protocols to optimize the quality of actual measurements, both for research purposes and clinical practice.

Another improvement in measurements may come from a new theory underlying the development of measurement instruments. More and more often the Item Response Theory (IRT) [16] is replacing the Classical Test Theory (CTT), which splits observed scores in true scores and error variances. The advantages of instruments developed according to the IRT are that their psychometric properties are less dependent on the population and situation, patients' scores can be compared if different versions of the tests are used, and the items of a test can be tailored to the severity of a patient's condition [5,16]. An important requirement for the application of the IRT is that the test is unidimensional [5,16]. Therefore, the value of the IRT for clinimetrics that focusses mainly of multidimensional indexes remains to be seen.

2.1.5. Clinimetrics in clinical practice

Although Feinstein defined clinimetrics as an important discipline for clinical practice, it has mainly been applied in clinical research, in which typically more attention is paid to measurement error than in clinical practice. However, considering the importance of measurement error, this should obviously be the other way around. In clinical research it is easier to detect measurement errors, because they can be seen on the scatter plot of repeated measurements, while in clinical practice there is usually only one measurement. Moreover, in clinical research it is easier to deal with measurement errors by looking at the averages for a group of persons, while in clinical practice there is often only one measurement per patient, so it is not possible to average any measurements. Therefore, the quality requirements for measurements in clinical practice should be higher than for clinical research. We believe that it is important to raise awareness about this issue within the medical community, and to offer practical solutions to improve the quality of measurements in clinical practice.

3. The future of clinimetrics

In modern medicine the importance of clinimetric issues is increasing for several reasons. First, the strong emphasis on evidence-based medicine requires the value of clinical interventions to be shown empirically, which implies measurement of outcomes. Second, because of the current high standards of medicine in the Western world, the added value of diagnostic tests and therapeutic interventions becomes smaller and smaller. So, if we still want to show improvements, our measurement instruments must meet the challenge to detect smaller changes. This will obviously increase the standards for the reproducibility and (longitudinal) validity or responsiveness of measurements. Third, it appears that we are currently encountering a number of "new" diseases that are rather challenging diagnostically, for example, repetitive strain injury and chronic fatigue syndrome. This especially emphasizes the need to further develop methods for the construct validation of diagnostic tests, typically by comparisons with indicators of symptom severity and prognosis.

Fourth, technical developments are continuously improving the diagnostic potentials, for example, by making more anatomic structures and pathophysiologic processes visible. But these data still have to be interpreted, and because the observations and interpretations often have to be made by the medical specialists, this keeps them dependent on the views and opinions of care providers. All new measurement instruments require a critical evaluation of their clinimetric properties and their relevance for diagnosis, prognosis, or choice of treatment. The emphasis of this evaluation should not only be on technical performance, but also on interpretation of the data.

Close collaboration between clinicians, statisticians, epidemiologists, and psychologists is necessary to guarantee a healthy future development of clinimetrics, serving the needs of both clinical research and clinical practice.

References

[1] Feinstein AR. An additional basic science for clinical medicine: IV The development of clinimetrics. Ann Intern Med 1983;99:843–8.
[2] Feinstein AR. Clinimetrics. New Haven, CT: Yale University Press; 1987.
[3] Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27–36.
[4] Guyatt G, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties. J Clin Epidemiol 1992;45:1341–5.
[5] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Press; 1995.
[6] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40: 171–8.
[7] De Vet HCW, Beurskens AJHM, Bezemer PD, Bouter LM. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Health Technol Assess Health Care 2001;17:479–87.
[8] De Vet HCW. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopedia biostatistica. vol. 4. Chicester: John Wiley & Sons, Ltd.; 1998, p. 3123–8.

[9] Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12:349–62.
[10] Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001;54:1204–17.
[11] Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459–68.
[12] Norman GR, Gwadry Sridhar F, Guyatt G, Walter SD. Relation of distribution- and anchor based approaches in interpretation of changes in health related quality of life. Med Care 2001;39:1039–47.
[13] Bialocerkowski AE, Grimmer KA, Bain GI. A systematic review of the content and quality of wrist outcome instruments. Int J Qual Health Care 2000;12:149–57.
[14] Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality of life instruments. Pharmacoeconomics 2000;17: 13–35.
[15] McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care 2000;38(suppl II):43–59.
[16] Hambleton RK, Swaminathan H. Item response theory: principles and applications. Boston: Kluwer Nijhoff; 1985.